J. Clin. Chem. Clin. Biochem.
Vol. 20, 1982, 75–80

Determination of n-Dimensional Reference Ellipsoids Using Patient Data

By *A. J. Naus, A. Borst* and *P. S. Kuppens*

Department of Clinical Chemistry, St. Laurentius Hospital, Roermond, The Netherlands

(Received July 27/November 11, 1981)

Summary: A method is described for the calculation of n-dimensional reference ellipsoids, using patient data. The advantages and drawbacks of the use of reference ellipsoids for a set of different parameters, in contrast with the use of a reference range for every single parameter, are discussed. The use of reference ellipsoids in practice is illustrated with an example.

Bestimmung n-dimensionaler Referenzellipsoide mit Patienten-Daten

Zusammenfassung: Eine Methode für die Berechnung von n-dimensionalen Referenzellipsoiden aus Patienten-Daten wird beschrieben. Vorzüge und Nachteile der Verwendung von Referenzellipsoiden für einen Satz verschiedener Kenngrößen im Gegensatz zur Verwendung eines Referenzbereichs für jede einzelne Kenngröße werden erörtert. Die Verwendung von Referenzellipsoiden in der Praxis wird an einem Beispiel illustriert.

Introduction

In clinical chemistry it is customary to compare the result of an analysis with a reference range. In most cases this range is determined in such a way that 95% of a normal population lies within it. A result below the lower or above the upper limit is probably pathological, although the chance that it is normal is still 5% (1–4).

When in the same sample a second (independent) analyte is determined, the chance that the result of this second analysis is within its reference range, again is 95% (assuming that the person is healthy). The chance that both results are "normal" is $0.95^2 = 0.903$. In other words: the more independent analyses are performed in a sample of a healthy person, the greater the chance that one or more of the results are pathological. When 14 independent analyses are performed this chance is about 50% (5). Therefore, in order to be able to better differentiate between normal and pathological, it is advisable to use 2 or more dimensional reference ellipsoids rather than a reference range for every separate determination. A reference ellipsoid can be defined as the area in the n-dimensional space, where the chance that a set of n results of a healthy person lies, is 95%.

In theory it is possible to calculate a reference ellipsoid for every combination of n determinations. In practice, however, it is very difficult to use reference ellipsoids with a dimension greater than 2 without the use of a computer. Even when a computer can be used, it should be emphasized that the danger exists that a set of n results is indeed classified as pathological, but the reason for this classification is no longer apparent. Therefore in our view the use of reference ellipsoids should be limited to dimensions 2 and 3.

Materials and Methods

The general equation for a reference ellipsoid in the k-dimensional space for k variables that all have a *Gauss*ian frequency distribution, is given by:

$$S = \{Y | (Y - X)^T V^{-1} (Y - X) \leq \chi^2(\alpha, k)\} \qquad \text{eq. 1}$$

where:

S = reference ellipsoid,
Y = vector of k results,
X = vector of the means of the k determinations,
V = variance-covariance matrix,
k = dimension,
$\chi^2(\alpha, k)$ = α-fractile of a χ^2 distribution with k degrees of freedom.

When k = 1, in other words, when a reference range is calculated for a single variable, equation 1 simplifies to:

$$S = \{Y | (y - \mu)\ (1/\sigma^2)\ (y - \mu) \leq 3.84\} \qquad \text{eq. 2}$$

where:

μ = mean of the variable,
σ = standard deviation of the variable,
$\chi^2(0.95, 1) = 3.84$.

Furthermore, when $\mu = 0$ and $\sigma = 1$ (standard normal distribution), equation 2 becomes:

$$-1.96 <= Y <= 1.96 \qquad \text{eq. 3}$$

0340-076X/82/0020-0075$02.00

Equation 3 is of course very often used, when the reference range for a single parameter has to be calculated.

When equation 1 is used for a combination of 2 variables, the result is a circle when the coefficient of correlation is 0, or an ellipse when $|r| > 0$. As $|r|$ increases, the ellipse becomes slimmer (fig. 1). When the means of both variables are 0 and the variances 1, the variance-covariance matrix is given by:

$$V = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \quad \text{or} \quad V^{-1} = \frac{1}{1-r^2} \cdot \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}. \qquad \text{eq. 4}$$

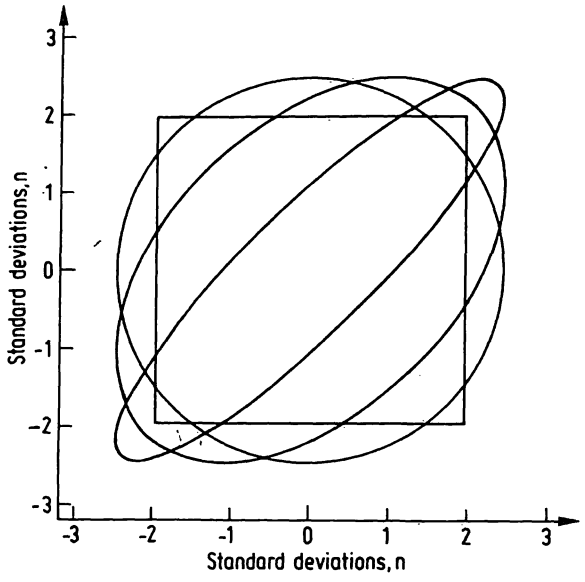


Fig. 1. Reference ellipsoids for a combination of two determinations ($\mu_1 = \mu_2 = 0$; $\sigma_1 = \sigma_2 = 1$) for r = 0, 0.45 and 0.90. The indicated square is the reference area used in the conventional way.

The tabulated value for $\chi^2(0.95, 2)$ is 5.99. The construction of figure 1 is now straightforward. In this figure, the square indicates the reference area used in the conventional way for a combination of these two determinations. It is clear that certain results lie within the square but outside the reference ellipse, in other words they are normal in the conventional way, but pathological with respect to the reference ellipsoid; the reverse can also occur.

In haematology it is customary to calculate from a set of values[1]) for Hb, RBC and PCV, the so called *Wintrobe* indices[1]) MCV, MCH and MCHC (7). These indices have various uses; for example, the combination of a high normal value for Hb and a low normal value for RBC is pathological, although the two results are each within their respective reference ranges. In this example, however, the calculated value for MCH would result in a pathological value, as should be the case. When, in the 3-dimensional space, the area is drawn within which Hb, RBC, PCV and MCV, MCH, MCHC are "normal", the result is a kind of prisma, which closely resembles an ellipsoid.

From the above equations it is clear that the most important thing to determine when a reference ellipsoid has to be calculated, is the variance-covariance matrix. This question can be resolved in two separate problems:

1. the determination of the variances of the separate variables and

2. the determination of the covariance of every combination of two variables.

The mean and variance of an assay can be determined by analysing a group of normal persons. The problems however in finding such a group are numerous. It certainly is not acceptable in our view to use the laboratory staff or a group of blood donors for this purpose, simply because they do not form a true representation of the whole population, although they may all be "normal". Automatically a selection is made when choosing one of these groups for the calculation of reference ranges (8, 9). The danger that these ranges are biased when a selection is made beforehand is very great, so in our view it is better to make no selection at all.

Simply take all the results produced during a certain time in your laboratory and use these. Of course a number of these results is "abnormal" and should not be used for the calculation of mean and standard deviation. In practice, however, most results for a routine test are completely normal. The *Bhattacharya* plot (10–13) is a statistical method that insures that the abnormal results in a frequency distribution do not influence the calculation of mean and standard deviation.

The *Bhattacharya* plot is based on the following:

the results of a determination are accumulated in equally spaced classes. If the frequency distribution is *Gaussian*, the logarithm of the quotient of the frequencies in class (i + 1) and class i, plotted against the midpoint of class i, results in a straight line. The mean and standard deviation of the distribution can be calculated from the x-intercept and the slope of the straight line respectively.

$$\mu = x_{\text{intercept}} - \frac{1}{2}h$$

$$\sigma^2 = -h/\text{slope} - h^2/12 \qquad \text{eq. 5}$$

where:

h = width of the classes.
$h^2/12$ = *Sheppard*'s correction for the grouping of data in classes.

When the number of test results, used to construct the *Bhattacharya* plot is small (less than 1500) the observed frequencies in every class should be smoothed. The method of choice for doing this, is the method of *Savitzky* et al. (14, 15).

So the *Bhattacharya* plot can result in values of μ and σ for a certain assay using unselected patient data. This means that the diagonal elements of the variance-covariance matrix can be determined quite easily. The remaining problem is to calculate values for the covariances or, since $\text{cov}(x, y) = r * \sigma(x) * \sigma(y)$, values for the coefficient of correlation when results for assay X are plotted against those of assay Y. When the frequency distribution of both determination X and determination Y is *Gaussian*, then the frequency distribution of a linear combination of X and Y is also *Gaussian*. This means that, when the *Bhattacharya* plot can be applied to determination $X(\mu_x, \sigma_x)$ and determination $Y(\mu_y, \sigma_y)$, this plot can also be applied to the sum of the results of the two determinations or the difference. When the sum of the results is used the mean is equal to $\mu_x + \mu_y$ and the variance is equal to $\sigma_x^2 + \sigma_y^2 + 2\,\text{cov}_{x,y}$; when the difference is used, the mean is $\mu_x - \mu_y$ and the variance is equal to $\sigma_x^2 + \sigma_y^2 - 2\text{cov}_{x,y}$.

So the covariance for a certain combination of two determinations (both having a *Gaussian* frequency distribution) can be calculated by the following equation:

$$\text{cov}_{x,y} = \frac{\sigma^2_{\text{sum}} - \sigma^2_{\text{diff}}}{4} \qquad \text{eq. 6}$$

Another method that can be used to determine the covariance is the following:

– Starting with a value for r equal 0, all data points are selected that lie within the 99% circle.

– Using the method of least squares a straight line is calculated through these points, resulting in a new value for r.

[1]) Hb = haemoglobin
RBC = red blood corpuscles, erythrocytes
PCV = packed corpuscular volume, haematocrit
MCV = mean corpuscular volume
MCH = mean corpuscular haemoglobin
MCHC = mean corpuscular haemoglobin concentration

– With this new r, again a selection of data points is made; the selection criterion is now that all points lie within the 99% ellipse.

– A straight line is again calculated through these points, resulting in a value for r that is slightly different from the old one.

– This process is repeated until the value of r no longer changes. With this r the covariance between determination X and Y is calculated.

This procedure is depicted in figure 2.

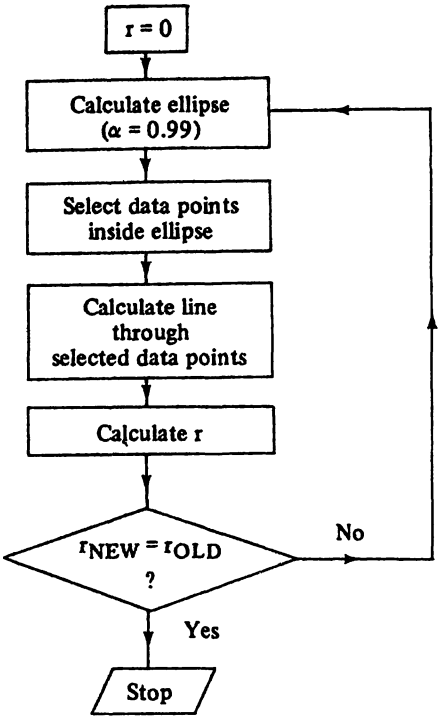


Fig. 2. Flow diagram for the calculation of the coefficient of correlation r between 2 determinations, using patient data.

Tab. 1. Calculation of the *Bhattacharya* plot for total protein.

i	Class mean	Observed frequency	Smoothed frequency	$\log(\frac{f_{i+1}}{f_i})$
1	41	2	–	–
2	43	4	–	–
3	45	2	2.4	- 0.1521
4	47	2	2.0	1.0028
5	49	5	5.6	0.4997
6	51	11	9.3	0.2543
7	53	12	12.0	0.5592
8	55	19	21.1	0.6985
9	57	42	42.5	0.2871
10	59	66	56.6	0.4026
11	61	70	84.7	0.3875
12	63	142	124.8	0.5536
13	65	198	217.2	0.4789
14	67	354	350.6	0.3410
15	69	502	493.1	0.1676
16	71	573	583.2	0.0109
17	73	601	589.6	- 0.1783
18	75	501	493.2	- 0.3844
19	77	321	335.8	- 0.5472
20	79	199	194.2	- 0.6387
21	81	100	102.5	- 0.8213
22	83	44	45.1	- 0.8632
23	85	21	19.0	- 0.7550
24	87	8	8.9	–
25	89	5	–	–
26	91	1	–	–

Tab. 2. Determination of covariance for total protein – albumin.

Determination	Mean	Variance
Total protein	71.9	23.72
Albumin	44.0	9.85
Total protein + albumin	117.5	45.40
Total protein - albumin	28.2	15.44

Results and Discussion

As an example we want to present the calculation of a reference ellipse for the combination total protein and albumin. The results for both determinations are produced by the SMA 12/60. This analyser is coupled to a teletype which punches all results in paper tape. This punched tape is then fed into a Wang PCS II desk top computer.

The frequency distributions for total protein, albumin, total protein + albumin and total protein - albumin are given in figure 3a–d (3840 results for every determination). After applying a 5 point quadratic smooth according to *Savitzky* et al. (14, 15) the *Bhattacharya* plots were constructed (fig. 4a–d). As an example, the details of the calculation of the *Bhattacharya* plot for albumin are given in table 1.

The resulting means and variances are summarized in table 2. From this table it follows that the coefficient of correlation (r) between total protein and albumin is 0.49.

Using the method depicted in figure 2, gives a value for r of 0.52, which is in close agreement.

With these results the 95% reference ellipse can easily be determined as:

$$S = \{Y | (x - 71.9 \;\; y - 44.0) \begin{pmatrix} 4.87^2 & 7.49 \\ 7.49 & 3.14^2 \end{pmatrix} \begin{pmatrix} x - 71.9 \\ y - 44.0 \end{pmatrix} \leqslant 5.99\} \qquad \text{eq. 7}$$

From figure 5 it can be seen that the number of data points that lie within the ellipse, but are abnormal in the conventional way, for total protein, albumin or both, is considerable (n = 153, i.e. 4.0%). More interesting is the rather great number of data points that are normal in the conventional way for total protein and albumin, but lie outside the reference ellipse (n = 43, i.e. 1.1%).

These data points are combinations of low normal total protein and high normal albumin or high normal total protein and low normal albumin, the combination of which is abnormal.

The method, described in this paper, for the calculation of n-dimensional reference ellipsoids, can only be applied when the frequency distribution of every determination is *Gaussian*. When, in the *Bhattacharya* plot, a sufficiently

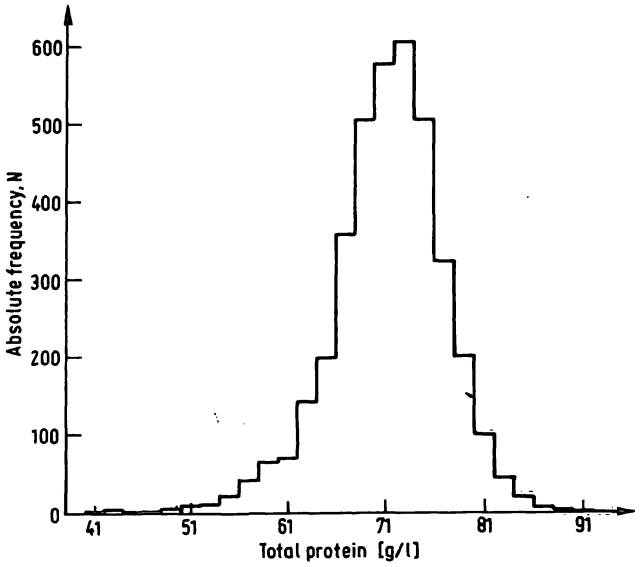


Fig. 3a

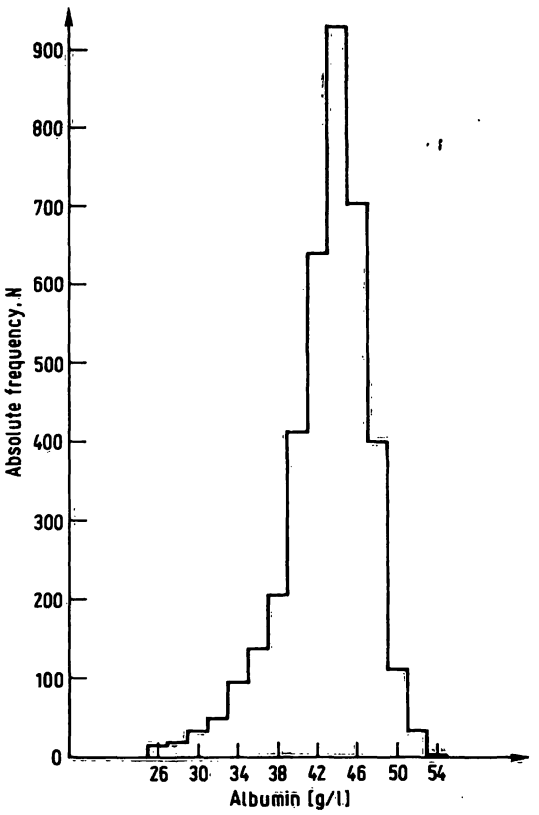


Fig. 3b

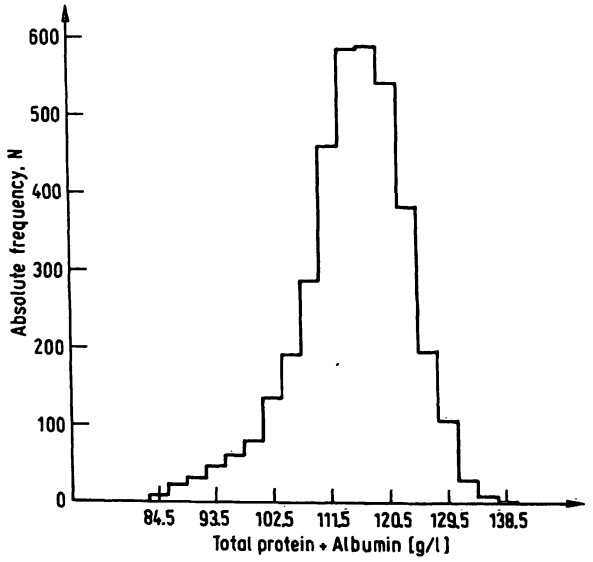


Fig. 3c

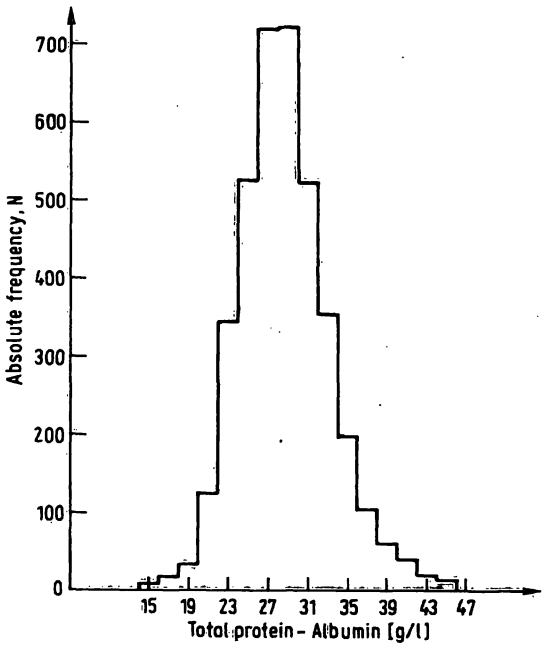


Fig. 3d

Fig. 3. Frequency distribution of 3840 values for total protein, albumin, total protein + albumin and total protein – albumin.

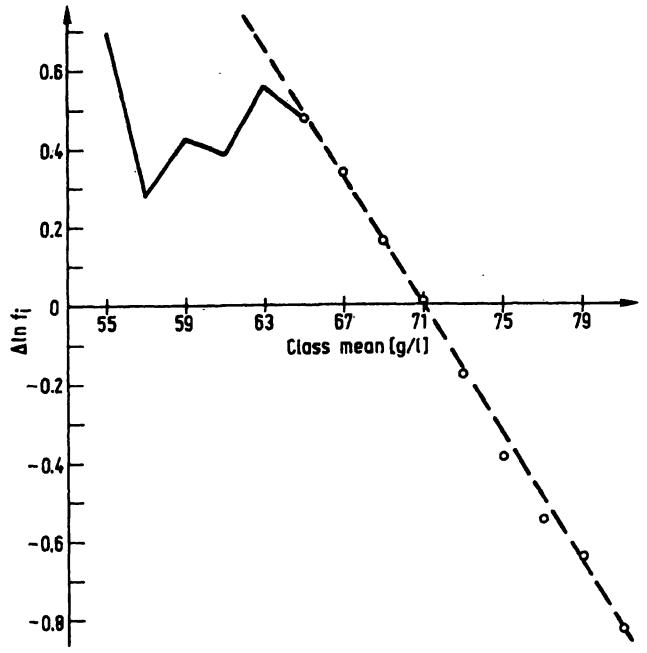


Fig. 4a

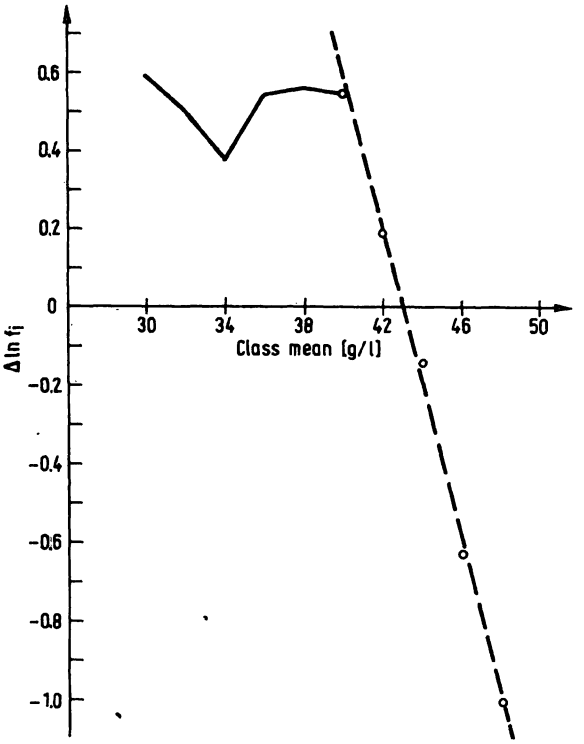


Fig. 4b

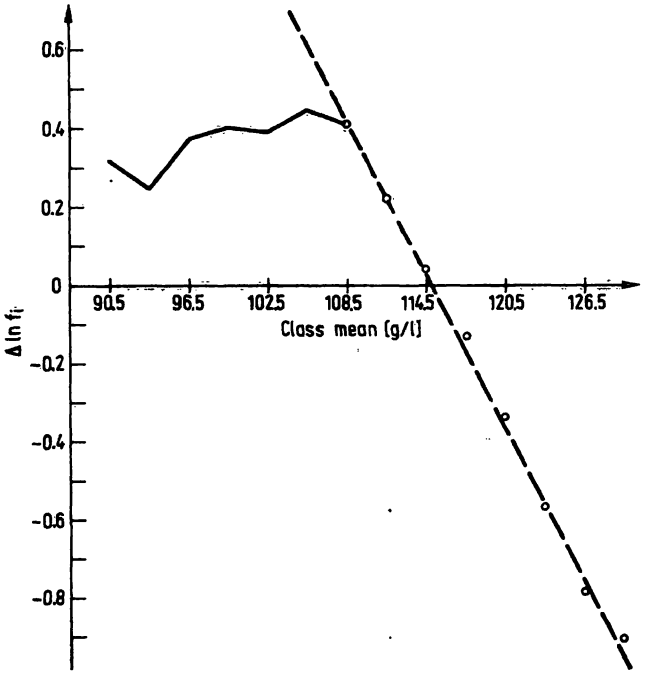


Fig. 4c

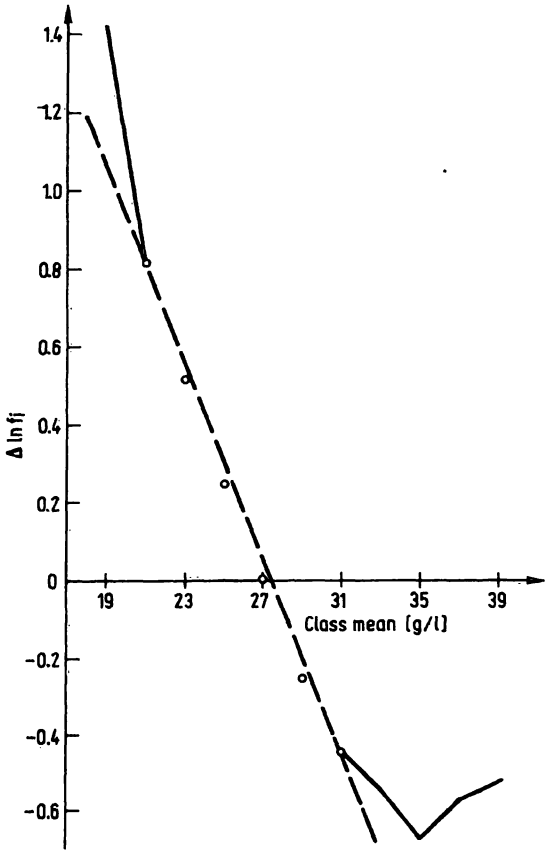
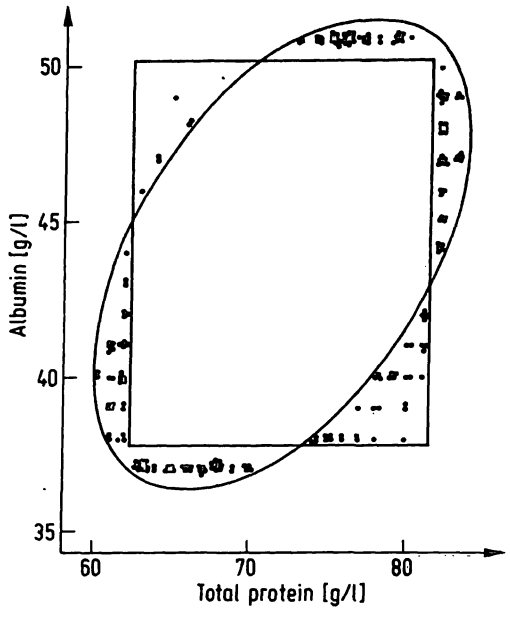


Fig. 4d

Fig. 4. *Bhattacharya* plots constructed with the data of fig. 3 after applying a 5 point quadratic smooth.



long straight part can be detected, the assumption that the frequency distribution is *Gaussian* can safely be made. The question, how to proceed, when the results of a determination are not distributed according to a *Gauss* function, is subject of further research, the results of which will be reported in a subsequent paper.

In our view reference ellipsoids can only be applied efficiently when a computer is used, especially when the dimension of the ellipsoid is greater than 2. Reference ellipsoids can hardly be printed in reference booklets for use by the clinician. This fact is an important drawback for the practical application of this undoubtedly very useful procedure.

Fig. 5. Calculated reference ellipse for the combination total protein albumin, using the data of fig. 3. By the conventional procedure the percentage of normal results is 2.9% less than in the ellipse method.

References

1. Rümke, C. & Bezemer, P. (1972) Ned. Tijdschr. Geneeskd. *116*, 1224–1230.
2. Rümke, C. & Bezemer, P. (1972) Ned. Tijdschr. Geneeskd. *116*, 1559–1568.
3. Dybkaer, R. & Gräsbeck, R. (1973) Scand. J. Clin. Lab. Invest. *32*, 1–7.
4. Dybkaer, R., Jörgensen, K. & Nyboe, J. (1975) Scand. J. Clin. Lab. Invest. *35*, Suppl. 144, 45–74.
5. Gross, R. & Oette, K. (1980) Die Problematik der ungezielten Mehrfachanalyse aus der Sicht des Klinikers. Technilab nov. 1980.
6. Guttman, I. (1970) Statistical tolerance regions. In: Griffin's statistical monographs & courses nr. 26. Griffin, London.
7. Wintrobe, M. (1974) Clinical hematology, 7th ed., Lea & Febiger, Philadelphia.
8. Hoeke, J. (1979) Normale waarden omstreden. In: Het medisch jaar 1979, 420–429, Bohn, Scheltema & Holkema, Utrecht.
9. Alström, T., Gräsbeck, R., Hjelm, M. & Skandsen, S. (1975) Scand. J. Clin. Lab. Invest. *35*, suppl. 144, 1–44.
10. Bhattacharya, C. (1967) Biometrics *23*, 115–135.
11. Naus, A., Borst, A. & Kuppens, P. (1980) J. Clin. Chem. Clin. Biochem. *18*, 621–625.
12. Gindler, E. (1970) Clin. Chem. *16* 124–128.
13. White, J. (1978) Clin. Chim. Acta *84*, 353–360.
14. Savitzky, A. & Golay, M. (1964) Anal. Chem. *36*, 1627–1638.
15. Steinier, J., Termonia, Y. & Deltour, J. (1972) Anal. Chem. *44*, 1906–1909.

Ir. A. J. Naus
Dept. of Clin. Chemistry
St. Laurentiushospital
Mgr. Driessenstraat 6
NL-6043 CV Roermond